

**Method, Computer Program and Data Processing System for Data  
Clustering**

5

**BACKGROUND OF THE INVENTION**

**Field of Invention**

The present invention relates to the field of data clustering and in particular to clustering algorithms and quality determination.

10

**Description of the Related Art**

15 Clustering of data is a data processing task in which clusters are identified in a structured set of raw data. Typically, the raw data consists of a large set of records, each record having the same or a similar format. Each field in a record can take any of a number of logical, categorical, or 20 numerical values. Data clustering aims to group such records into clusters such that records belonging to the same cluster have a high degree of similarity.

20

A variety of algorithms is known for data clustering. The 25 K-means algorithm relies on the minimal sum of Euclidean distances to centers of clusters, taking into consideration the number of clusters. The Kohonen algorithm is based on a neural net and also uses Euclidean distances. IBM's demographic algorithm relies on the sum of internal similarities minus the 30 sum of external similarities as a clustering criterion. Those

and other clustering criteria are utilized in an iterative process of finding clusters.

A common disadvantage of such prior art clustering algorithms is that different clustering algorithms applied to the same set of data may deliver largely different results. Even if the same algorithm is applied to the same set of data using a different set of parameters as a starting condition, a different result is likely to occur. In the prior art, no objective criterion exists to compare the results of such clustering operations.

One field of application of data clustering is data mining. US Patent No. 6,112,194 describes a technique for data mining including a feedback mechanism for monitoring performance of mining tasks. A user-selected mining technique type is received for the data mining operation. A quality measure type is identified for the user-selected mining technique type. The user-selected mining technique type for the data mining operation is processed and a quality indicator is measured using the quality measure type. The measured quality indication is displayed while processing the user-selected mining technique type for the data mining operations.

US Patent No. 6,115,708 describes a method for refining the initial conditions for clustering with applications to small and large database clustering. How this method is applied to the popular K-means clustering algorithm and how refined initial starting points indeed lead to improved solutions are described. The technique can be used as an initializer for

other clustering solutions. The method is based on an efficient technique for estimating the modes of a distribution and runs in time guaranteed to be less than overall clustering time for large data sets. The method is also scalable and hence can be  
5 efficiently used on huge databases to refine starting points for scalable clustering algorithms in data mining applications.

US Patent No. 6,100,901 describes a method for visualizing a multi-dimensional data set in which the multi-dimensional  
10 data set is clustered into k clusters, with each cluster having a centroid. Either two distinct current centroids or three distinct non-collinear current centroids are selected. A current 2-dimensional cluster projection is generated based on the selected current centroids. In the case when two distinct current centroids are selected, two distinct target centroids are selected, with at least one of the two target centroids being different from the two current centroids.

US Patent No. 5,857,179 describes a computer-implemented  
20 technique for clustering documents and automatic generation of cluster keywords. An initial document by term matrix is formed, each document being represented by a respective M dimensional vector, where M represents the number of terms or words in a predetermined domain of documents. The dimensionality of the  
25 initial matrix is reduced to form resultant vectors of the documents. The resultant vectors are then clustered such that correlated documents are grouped into respective clusters. For each cluster, the terms having greatest impact on the documents in that cluster are identified. The identified terms represent  
30 key words of each document in that cluster. Further, the

identified terms form a cluster summary indicative of the documents in that cluster.

#### SUMMARY OF THE INVENTION

5

A principal object of the present invention is to provide a method, data processing system and computer program product for data clustering and quality determination such that the qualities of clustering results can be compared on an objective basis. The quality index for a clustering result obtained in accordance with the invention is independent of the clustering algorithm used.

Rather than relying on the clustering algorithm itself for quality determination, the invention relies on a statistical analysis of the clustering result to determine the quality of the clustering. The statistical analysis uses a comparison of the foreground and background frequencies of buckets. The comparison results in a statistical parameter used to calculate a quality index.

According to a preferred embodiment, the quality index is normalized such that even if different sets of data are used as a basis for different clustering operations, the results of the clustering are still comparable based on the objective quality index.

According to a further preferred embodiment of the invention, a clustering operation is carried out by performing a data clustering operation based on a variety of different

clustering algorithms either in parallel or sequentially, determining the qualities of the respective clustering results and ranking the results accordingly. The result with the highest quality index can be considered the overall result of  
5 the clustering operation.

Further, the invention provides a clustering algorithm relying on an objective quality index to be optimized in a number of iterations. This algorithm outputs a resulting  
10 quality index for its clustering result which is objective and can be compared to corresponding other results.

A method of the invention is advantageously implemented in a data processing system by means of a corresponding computer program. If a number of different clustering algorithms is used, it is advantageous to assign a dedicated processing unit of the data processing system to each clustering algorithm for the purpose of parallel processing. This has the advantage of minimizing the processing time required.  
15  
20

#### BRIEF DESCRIPTION OF THE DRAWINGS

The present invention together with the above and other objects and advantages may best be understood from the following description of the preferred embodiments of the invention as illustrated in the drawings, wherein:  
25  
30

Fig. 1 is a schematic representation of the structure of a cluster j;

Fig. 2 is a flow chart illustrating a preferred embodiment of the determination of a quality index;

5 Fig. 3 is a flow chart illustrating the utilization of different clustering algorithms in parallel;

10 Fig. 4 is a flow chart illustrating a clustering algorithm relying on an objective criterion to be optimized in a number of iterations; and

15 Fig. 5 is a block diagram showing the structure of a data processing system.

20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100  
105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200  
205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300  
305 310 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390 395 400  
405 410 415 420 425 430 435 440 445 450 455 460 465 470 475 480 485 490 495 500  
505 510 515 520 525 530 535 540 545 550 555 560 565 570 575 580 585 590 595 600  
605 610 615 620 625 630 635 640 645 650 655 660 665 670 675 680 685 690 695 700  
705 710 715 720 725 730 735 740 745 750 755 760 765 770 775 780 785 790 795 800  
805 810 815 820 825 830 835 840 845 850 855 860 865 870 875 880 885 890 895 900  
905 910 915 920 925 930 935 940 945 950 955 960 965 970 975 980 985 990 995 1000  
1005 1010 1015 1020 1025 1030 1035 1040 1045 1050 1055 1060 1065 1070 1075 1080 1085 1090 1095 1100  
1105 1110 1115 1120 1125 1130 1135 1140 1145 1150 1155 1160 1165 1170 1175 1180 1185 1190 1195 1200  
1205 1210 1215 1220 1225 1230 1235 1240 1245 1250 1255 1260 1265 1270 1275 1280 1285 1290 1295 1300  
1305 1310 1315 1320 1325 1330 1335 1340 1345 1350 1355 1360 1365 1370 1375 1380 1385 1390 1395 1400  
1405 1410 1415 1420 1425 1430 1435 1440 1445 1450 1455 1460 1465 1470 1475 1480 1485 1490 1495 1500  
1505 1510 1515 1520 1525 1530 1535 1540 1545 1550 1555 1560 1565 1570 1575 1580 1585 1590 1595 1600  
1605 1610 1615 1620 1625 1630 1635 1640 1645 1650 1655 1660 1665 1670 1675 1680 1685 1690 1695 1700  
1705 1710 1715 1720 1725 1730 1735 1740 1745 1750 1755 1760 1765 1770 1775 1780 1785 1790 1795 1800  
1805 1810 1815 1820 1825 1830 1835 1840 1845 1850 1855 1860 1865 1870 1875 1880 1885 1890 1895 1900  
1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000  
2005 2010 2015 2020 2025 2030 2035 2040 2045 2050 2055 2060 2065 2070 2075 2080 2085 2090 2095 2100  
2105 2110 2115 2120 2125 2130 2135 2140 2145 2150 2155 2160 2165 2170 2175 2180 2185 2190 2195 2200  
2205 2210 2215 2220 2225 2230 2235 2240 2245 2250 2255 2260 2265 2270 2275 2280 2285 2290 2295 2300  
2305 2310 2315 2320 2325 2330 2335 2340 2345 2350 2355 2360 2365 2370 2375 2380 2385 2390 2395 2400  
2405 2410 2415 2420 2425 2430 2435 2440 2445 2450 2455 2460 2465 2470 2475 2480 2485 2490 2495 2500  
2505 2510 2515 2520 2525 2530 2535 2540 2545 2550 2555 2560 2565 2570 2575 2580 2585 2590 2595 2600  
2605 2610 2615 2620 2625 2630 2635 2640 2645 2650 2655 2660 2665 2670 2675 2680 2685 2690 2695 2700  
2705 2710 2715 2720 2725 2730 2735 2740 2745 2750 2755 2760 2765 2770 2775 2780 2785 2790 2795 2800  
2805 2810 2815 2820 2825 2830 2835 2840 2845 2850 2855 2860 2865 2870 2875 2880 2885 2890 2895 2900  
2905 2910 2915 2920 2925 2930 2935 2940 2945 2950 2955 2960 2965 2970 2975 2980 2985 2990 2995 3000  
3005 3010 3015 3020 3025 3030 3035 3040 3045 3050 3055 3060 3065 3070 3075 3080 3085 3090 3095 3100  
3105 3110 3115 3120 3125 3130 3135 3140 3145 3150 3155 3160 3165 3170 3175 3180 3185 3190 3195 3200  
3205 3210 3215 3220 3225 3230 3235 3240 3245 3250 3255 3260 3265 3270 3275 3280 3285 3290 3295 3300  
3305 3310 3315 3320 3325 3330 3335 3340 3345 3350 3355 3360 3365 3370 3375 3380 3385 3390 3395 3400  
3405 3410 3415 3420 3425 3430 3435 3440 3445 3450 3455 3460 3465 3470 3475 3480 3485 3490 3495 3500  
3505 3510 3515 3520 3525 3530 3535 3540 3545 3550 3555 3560 3565 3570 3575 3580 3585 3590 3595 3600  
3605 3610 3615 3620 3625 3630 3635 3640 3645 3650 3655 3660 3665 3670 3675 3680 3685 3690 3695 3700  
3705 3710 3715 3720 3725 3730 3735 3740 3745 3750 3755 3760 3765 3770 3775 3780 3785 3790 3795 3800  
3805 3810 3815 3820 3825 3830 3835 3840 3845 3850 3855 3860 3865 3870 3875 3880 3885 3890 3895 3900  
3905 3910 3915 3920 3925 3930 3935 3940 3945 3950 3955 3960 3965 3970 3975 3980 3985 3990 3995 4000  
4005 4010 4015 4020 4025 4030 4035 4040 4045 4050 4055 4060 4065 4070 4075 4080 4085 4090 4095 4100  
4105 4110 4115 4120 4125 4130 4135 4140 4145 4150 4155 4160 4165 4170 4175 4180 4185 4190 4195 4200  
4205 4210 4215 4220 4225 4230 4235 4240 4245 4250 4255 4260 4265 4270 4275 4280 4285 4290 4295 4300  
4305 4310 4315 4320 4325 4330 4335 4340 4345 4350 4355 4360 4365 4370 4375 4380 4385 4390 4395 4400  
4405 4410 4415 4420 4425 4430 4435 4440 4445 4450 4455 4460 4465 4470 4475 4480 4485 4490 4495 4500  
4505 4510 4515 4520 4525 4530 4535 4540 4545 4550 4555 4560 4565 4570 4575 4580 4585 4590 4595 4600  
4605 4610 4615 4620 4625 4630 4635 4640 4645 4650 4655 4660 4665 4670 4675 4680 4685 4690 4695 4700  
4705 4710 4715 4720 4725 4730 4735 4740 4745 4750 4755 4760 4765 4770 4775 4780 4785 4790 4795 4800  
4805 4810 4815 4820 4825 4830 4835 4840 4845 4850 4855 4860 4865 4870 4875 4880 4885 4890 4895 4900  
4905 4910 4915 4920 4925 4930 4935 4940 4945 4950 4955 4960 4965 4970 4975 4980 4985 4990 4995 5000  
5005 5010 5015 5020 5025 5030 5035 5040 5045 5050 5055 5060 5065 5070 5075 5080 5085 5090 5095 5100  
5105 5110 5115 5120 5125 5130 5135 5140 5145 5150 5155 5160 5165 5170 5175 5180 5185 5190 5195 5200  
5205 5210 5215 5220 5225 5230 5235 5240 5245 5250 5255 5260 5265 5270 5275 5280 5285 5290 5295 5300  
5305 5310 5315 5320 5325 5330 5335 5340 5345 5350 5355 5360 5365 5370 5375 5380 5385 5390 5395 5400  
5405 5410 5415 5420 5425 5430 5435 5440 5445 5450 5455 5460 5465 5470 5475 5480 5485 5490 5495 5500  
5505 5510 5515 5520 5525 5530 5535 5540 5545 5550 5555 5560 5565 5570 5575 5580 5585 5590 5595 5600  
5605 5610 5615 5620 5625 5630 5635 5640 5645 5650 5655 5660 5665 5670 5675 5680 5685 5690 5695 5700  
5705 5710 5715 5720 5725 5730 5735 5740 5745 5750 5755 5760 5765 5770 5775 5780 5785 5790 5795 5800  
5805 5810 5815 5820 5825 5830 5835 5840 5845 5850 5855 5860 5865 5870 5875 5880 5885 5890 5895 5900  
5905 5910 5915 5920 5925 5930 5935 5940 5945 5950 5955 5960 5965 5970 5975 5980 5985 5990 5995 6000  
6005 6010 6015 6020 6025 6030 6035 6040 6045 6050 6055 6060 6065 6070 6075 6080 6085 6090 6095 6100  
6105 6110 6115 6120 6125 6130 6135 6140 6145 6150 6155 6160 6165 6170 6175 6180 6185 6190 6195 6200  
6205 6210 6215 6220 6225 6230 6235 6240 6245 6250 6255 6260 6265 6270 6275 6280 6285 6290 6295 6300  
6305 6310 6315 6320 6325 6330 6335 6340 6345 6350 6355 6360 6365 6370 6375 6380 6385 6390 6395 6400  
6405 6410 6415 6420 6425 6430 6435 6440 6445 6450 6455 6460 6465 6470 6475 6480 6485 6490 6495 6500  
6505 6510 6515 6520 6525 6530 6535 6540 6545 6550 6555 6560 6565 6570 6575 6580 6585 6590 6595 6600  
6605 6610 6615 6620 6625 6630 6635 6640 6645 6650 6655 6660 6665 6670 6675 6680 6685 6690 6695 6700  
6705 6710 6715 6720 6725 6730 6735 6740 6745 6750 6755 6760 6765 6770 6775 6780 6785 6790 6795 6800  
6805 6810 6815 6820 6825 6830 6835 6840 6845 6850 6855 6860 6865 6870 6875 6880 6885 6890 6895 6900  
6905 6910 6915 6920 6925 6930 6935 6940 6945 6950 6955 6960 6965 6970 6975 6980 6985 6990 6995 7000  
6995 7000 7005 7010 7015 7020 7025 7030 7035 7040 7045 7050 7055 7060 7065 7070 7075 7080 7085 7090  
7095 7100 7105 7110 7115 7120 7125 7130 7135 7140 7145 7150 7155 7160 7165 7170 7175 7180 7185 7190  
7195 7200 7205 7210 7215 7220 7225 7230 7235 7240 7245 7250 7255 7260 7265 7270 7275 7280 7285 7290  
7295 7300 7305 7310 7315 7320 7325 7330 7335 7340 7345 7350 7355 7360 7365 7370 7375 7380 7385 7390  
7395 7400 7405 7410 7415 7420 7425 7430 7435 7440 7445 7450 7455 7460 7465 7470 7475 7480 7485 7490  
7495 7500 7505 7510 7515 7520 7525 7530 7535 7540 7545 7550 7555 7560 7565 7570 7575 7580 7585 7590  
7595 7600 7605 7610 7615 7620 7625 7630 7635 7640 7645 7650 7655 7660 7665 7670 7675 7680 7685 7690  
7695 7700 7705 7710 7715 7720 7725 7730 7735 7740 7745 7750 7755 7760 7765 7770 7775 7780 7785 7790  
7795 7800 7805 7810 7815 7820 7825 7830 7835 7840 7845 7850 7855 7860 7865 7870 7875 7880 7885 7890  
7895 7900 7905 7910 7915 7920 7925 7930 7935 7940 7945 7950 7955 7960 7965 7970 7975 7980 7985 7990  
7995 8000 8005 8010 8015 8020 8025 8030 8035 8040 8045 8050 8055 8060 8065 8070 8075 8080 8085 8090  
8095 8100 8105 8110 8115 8120 8125 8130 8135 8140 8145 8150 8155 8160 8165 8170 8175 8180 8185 8190  
8195 8200 8205 8210 8215 8220 8225 8230 8235 8240 8245 8250 8255 8260 8265 8270 8275 8280 8285 8290  
8295 8300 8305 8310 8315 8320 8325 8330 8335 8340 8345 8350 8355 8360 8365 8370 8375 8380 8385 8390  
8395 8400 8405 8410 8415 8420 8425 8430 8435 8440 8445 8450 8455 8460 8465 8470 8475 8480 8485 8490  
8495 8500 8505 8510 8515 8520 8525 8530 8535 8540 8545 8550 8555 8560 8565 8570 8575 8580 8585 8590  
8595 8600 8605 8610 8615 8620 8625 8630 8635 8640 8645 8650 8655 8660 8665 8670 8675 8680 8685 8690  
8695 8700 8705 8710 8715 8720 8725 8730 8735 8740 8745 8750 8755 8760 8765 8770 8775 8780 8785 8790  
8795 8800 8805 8810 8815 8820 8825 8830 8835 8840 8845 8850 8855 8860 8865 8870 8875 8880 8885 8890  
8895 8900 8905 8910 8915 8920 8925 8930 8935 8940 8945 8950 8955 8960 8965 8970 8975 8980 8985 8990  
8995 9000 9005 9010 9015 9020 9025 9030 9035 9040 9045 9050 9055 9060 9065 9070 9075 9080 9085 9090  
9095 9100 9105 9110 9115 9120 9125 9130 9135 9140 9145 9150 9155 9160 9165 9170 9175 9180 9185 9190  
9195 9200 9205 9210 9215 9220 9225 9230 9235 9240 9245 9250 9255 9260 9265 9270 9275 9280 9285 9290  
9295 9300 9305 9310 9315 9320 9325 9330 9335 9340 9345 9350 9355 9360 9365 9370 9375 9380 9385 9390  
9395 9400 9405 9410 9415 9420 9425 9430 9435 9440 9445 9450 9455 9460 9465 9470 9475 9480 9485 9490  
9495 9500 9505 9510 9515 9520 9525 9530 9535 9540 9545 9550 9555 9560 9565 9570 9575 9580 9585 9590  
9595 9600 9605 9610 9615 9620 9625 9630 9635 9640 9645 9650 9655 9660 9665 9670 9675 9680 9685 9690  
9695 9700 9705 9710 9715 9720 9725 9730 9735 9740 9745 9750 9755 9760 9765 9770 9775 9780 9785 9790  
9795 9800 9805 9810 9815 9820 9825 9830 9835 9840 9845 9850 9855 9860 9865 9870 9875 9880 9885 9890  
9895 9900 9905 9910 9915 9920 9925 9930 9935 9940 9945 9950 9955 9960 9965 9970 9975 9980 9985 9990  
9995 9995 10000

DESCRIPTION OF THE PREFERRED EMBODIMENT

The raw data on which the data clustering operation is applied consists of a large volume of such structured data

records. The result of a clustering operation yields a number  $k$  of clusters of which the cluster  $j$  is schematically depicted in the example of Fig. 1.

5       The variable  $l=2$  has the value A in the record  $R-j1$ . In other words, the bucket  $i=1$  for the variable  $l=2$  in the record  $R-j1$  equals A. Other than A, the variable  $l=2$  can also take values B or C, i.e., the bucket  $i=2$  is B and the bucket  $i=3$  for this variable  $l=2$  is C, respectively. For example, in the 10 record  $R-j3$  of the cluster  $j$ , the variable  $l=2$  has the bucket C( $i=3$ ), and in the record  $R-j4$  of the cluster  $j$ , the variable  $l=2$  has the bucket A again( $i=1$ ).  
15

With respect to Fig. 2, a preferred embodiment of a method 15 for determining a quality index for a clustering result is now explained in more detail. In Step 20, the relative foreground frequency of a bucket  $i$  of the variable  $l$  is determined for the cluster  $j$ . For example, the relative foreground frequency of the bucket  $i=1$  for the variable  $l=2$  in the cluster  $j$  of the 20 example shown in Fig. 1 is  $3/5$ , as the bucket  $i=1$  for this variable, which is A, occurs three times in the total of the five records contained in the cluster  $j$ .

In the next Step 21, the relative background frequency of 25 the bucket  $i$  of the variable  $l$  is determined for all clusters, i.e., for the entire set of records contained in the clustered data. In the example considered with respect to Fig. 1, this is done by determining the number of occurrences of the bucket  $i=1$  for the variable  $l=2$  in all records and dividing the absolute 30 number of occurrences by the number of all records.

In Step 22, a comparison value is determined to compare the relative foreground and background frequencies resulting from steps 20 and 21. The comparison can be performed by  
5 subtracting the relative foreground and background frequencies for a given bucket i of a given variable l. This is reflected in the following equation:

10 (1)  $f_{j,i,l} - v_{i,l}$

where  $f_{j,i,l}$  is the relative foreground frequency of the bucket i of the variable l in the cluster j and  $v_{i,l}$  is the relative background frequency of the bucket i of the variable l. This subtraction yields a parameter which is representative of the differentiation of the cluster j in comparison to all other clusters as far as the bucket i of the variable l is concerned.  
15 As the result of the subtraction can be negative, it is advantageous to either square the result:

20 (2)  $(f_{j,i,l} - v_{i,l})^2$

25 or to determine the absolute value of the result:

(3)  $|f_{j,i,l} - v_{i,l}|.$

In Step 23, these comparison values are determined and then added for all buckets  $i$  in all clusters  $j$  for a given variable  $l$  according to the following equation:

$$(4) \quad r_l = \sum_{j=1}^k \sum_{i=1}^m (f_{j,i,l} - v_{i,l})^2$$

The resulting parameter  $r$ , is multiplied with a factor in Step 24. The factor is determined in steps 25 and 26. In Step 25, the optimal number of clusters (optClust) is determined. For example, the optimal number of clusters can be defined to be equal to the maximum number of buckets of any of the variables. It is advantageous to set a threshold value for the optimal number of clusters in case one of the variables has a very large number of buckets or if the maximum number of clusters is dictated by the purpose of the clustering operation. For example, if the clustering is performed to identify demographic groups of people for group oriented advertisement typically not more than ten clusters corresponding to ten different marketing campaigns or segments are desirable.

In Step 26, the factor is calculated based on the optimal number of clusters and the actual number of clusters. The actual number of clusters is the number of clusters resulting from the clustering operation.

In Step 27, a division by the number of variables  $n$  is performed. The summation of the parameter  $r_i$  for all variables  $i$

yields the quality index QI according to the following equation:

$$(5) \quad QI = \frac{1}{n} * \sum_{l=1}^n r_l * \frac{\min[\text{optClust}, \text{NbrClust}]}{\max[\text{optClust}, \text{NbrClust}]}$$

5

where  $\min[\text{optClust}, \text{NbrClust}]$  is the smaller number of optClust and NbrClust and  $\max[\text{optClust}, \text{NbrClust}]$  is the bigger number.

The quality index QI is outputted in step 28.

According to a further preferred embodiment of the invention a normalizing value is determined to make the quality index independent of the data to which the clustering operation is applied. This has the advantage that even if clustering operations are performed on a different set of data, the quality of the results is still comparable. The normalizing value  $o_l$  for a given variable l is determined in accordance with the following equation:

$$(6) \quad o_l = \sum_{i=1}^m (1 - v_{i,l})^2 + (k-1) \sum_{i=1}^m (v_{i,l})^2$$

20

15

25

The equation 6 corresponds to the above equation 4 for the case of an imaginary situation where in one of the clusters the relative foreground frequency of a bucket is equal to one and equal to zero for all other clusters. In other words, All records containing the bucket are concentrated in the same cluster. This cluster corresponds to the first summation term in equation 6; all the other clusters are represented by the

second summation term multiplied by the number of clusters k minus 1.

This way the normalized quality index is determined in  
5 accordance with following equation:

$$(7) \quad QI = \frac{1}{n} * \sum_{l=1}^n \frac{o_l}{o_l} * \frac{\min[\text{optClust}, \text{NbrClust}]}{\max[\text{optClust}, \text{NbrClust}]}$$

Fig. 3 shows an example of an application of the method of  
10 Fig. 2 for performing a clustering of structured data 30 comprising records similar to the records of Fig. 1. The clustering algorithms CL 1, CL 2... CL q are applied on the data 30. This yields the clustering results RES 1, RES 2... RES q. For each of the results, a corresponding quality index QI 1, QI 2,...  
15 QI q is determined in accordance with the method of Fig. 2. This is done by means of parallel data processing in Steps 31, 32 and 33, respectively.

In Step 34, the quality indices QI 1, QU 2,... QU q are  
20 evaluated by numeric comparison. The numeric comparison of the quality indices results in an ordered list of the quality indices corresponding to a ranking of the respective results. The comparison of the quality of the results is made possible by the invention because it allows to determine an objective  
25 quality index for each result purely based on a statistical analysis of the result without relying on the clustering algorithm used to obtain the result.

The ranking of the result is outputted in Step 35. The result with the highest quality index QI can be considered the overall end result of the data clustering operation of Fig. 3.

5 With respect to Fig. 4, a clustering method being based on the objective quality index of the invention is shown in more detail. The clustering method is applied to a set of structured data 40 comprising records substantially similar to the example Fig. 1. In Step 41, a convenient initial set of clusters is  
10 selected. This can be done by using any of the known clustering methods. In Step 42, the quality index  $Q(\text{initial})$  for the initial set of clusters is calculated in accordance with equation (5) or (7).

15 In Step 43, the initial set of clusters is modified by moving one or more records from their clusters to other clusters. In Step 44, the quality index  $Q(\text{modified})$  for the modified set of clusters is calculated in accordance with equation (5) or (7).

20 In Step 45, it is decided whether the quality index  $Q(\text{modified})$  is greater than the quality index  $Q(\text{initial})$ . If this is not the case, this implies that the quality of the clustering did not improve. As a consequence, the modification  
25 previously performed in Step 43 is reversed in Step 46 and the control returns to Step 43 to perform a different modification.

30 In case the result of Step 45 is that in fact  $Q(\text{modified})$  is greater than  $Q(\text{initial})$  and thus the quality of the clustering increased, control of the process goes to Step 47.

In Step 47, it is decided if the actual number of iterations has been reached. If this is the case, the execution of the program stops in Step 48. If the contrary is the case, in Step 49 the modified set of clusters is declared to be the initial set of clusters for a further iteration step. This way the quality of the clustering is gradually increased until it reaches an ideal value or the operation is stopped after a predetermined number of iterations.

Fig. 5 shows a schematic block diagram of a preferred embodiment of a data processing system in accordance with the invention. The data processing system has a database 50 for storage of structured data. The database 50 is connected to a number of parallel processing units P1, P2, P3 and P4 via data bus 51. In each of the processing units P1 to P4, a data clustering operation is performed based on a variety of data clustering algorithms. The corresponding results are outputted to a control program stored in memory 52. The control program determines a quality index for each clustering result obtained by the parallel processing units P1 to P4. This is done in accordance with the preferred embodiments of Fig. 2 and Fig. 3. The clustering result with the highest quality index value is selected by the control program and outputted as result 53.